CHAPTER 1

Preliminaries

1. Least Squares Approximation

Let V be a vector space, with vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots$ and scalars $\alpha, \beta, \ldots$. The space V is an inner product space if one has defined a function $(\cdot, \cdot)$ from $V \times V$ to the reals (if the vector space is real) or to the complex (if V is complex) such that for all $\mathbf{u}, \mathbf{v} \in V$ and all scalars α the following conditions hold:

$$\begin{aligned} (\mathbf{u}, \mathbf{v}) &= \overline{(\mathbf{v}, \mathbf{u})}, \\ (\mathbf{u}, \mathbf{v} + \mathbf{w}) &= (\mathbf{u}, \mathbf{v}) + (\mathbf{u}, \mathbf{w}), \\ (\alpha\mathbf{u}, \mathbf{v}) &= \alpha(\mathbf{u}, \mathbf{v}), \\ (\mathbf{v}, \mathbf{v}) &\geq 0, \\ (\mathbf{v}, \mathbf{v}) = 0 &\Leftrightarrow \mathbf{v} = 0, \end{aligned} \tag{1.1}$$

where the overbar denotes the complex conjugate. Two elements $\mathbf{u}, \mathbf{v}$ such that $(\mathbf{u}, \mathbf{v}) = 0$ are said to be orthogonal.

The most familiar inner product space is $\mathbb{R}^n$ with the Euclidean inner product. If $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ then

$$(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} u_i v_i.$$

Another inner product space is $C[0, 1]$, the space of continuous functions on $[0, 1]$, with $(f, g) = \int_0^1 f(x)g(x)dx$.

The least squares, or "L_2" norm is

$$\|\mathbf{v}\| = \sqrt{(\mathbf{v}, \mathbf{v})}.$$

This has the following properties, which can be deduced from the properties of the inner product,

$$\begin{aligned} \|\alpha\mathbf{v}\| &= |\alpha|\|\mathbf{v}\| \\ \|\mathbf{v}\| &\geq 0, \\ \|\mathbf{v}\| = 0 &\Leftrightarrow \mathbf{v} = 0 \\ \|\mathbf{u} + \mathbf{v}\| &\leq \|\mathbf{u}\| + \|\mathbf{v}\|. \end{aligned}$$

The last, called the triangle inequality, follows from the Schwartz inequality

$$|(\mathbf{u}, \mathbf{v})| \leq \|\mathbf{u}\|\|\mathbf{v}\|.$$

In addition to these three properties, common to all norms, the L_2 norm has the "parallelogram property" (so-called because it is a property of parallelograms)

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2)$$

which can be verified by expanding the inner products.

Let $\{\mathbf{u}_n\}$ be a sequence in V.

DEFINITION. A sequence $\{\mathbf{u}_n\}$ is said to converge to $\hat{\mathbf{u}} \in V$ if $\|\mathbf{u}_n - \hat{\mathbf{u}}\| \to 0$ as $n \to \infty$ (i.e., for any $\epsilon > 0$ there exists some $N \in \mathbb{N}$ such that $n > N$ implies $\|\mathbf{u}_n - \hat{\mathbf{u}}\| < \epsilon$).

DEFINITION. A sequence $\{\mathbf{u}_n\}$ is a Cauchy sequence if given $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $m, n > N$ $\|\mathbf{u}_n - \mathbf{u}_m\| < \epsilon$.

A sequence which converges is a Cauchy sequence, although the converse is not necessarily true. If the converse is true for all Cauchy sequences in a given inner product space, then the space is called complete. We shall assume that all of the spaces we work with from now on are complete.

A few more definitions from real analysis:

DEFINITION. An open ball centered at $\mathbf{x}$ with radius $r > 0$ is the set $B_r(\mathbf{x}) = \{\mathbf{u} : \|\mathbf{u} - \mathbf{x}\| < r\}$.

DEFINITION. A set S is closed if for every $\mathbf{x} \in S$ there exists a sequence $\{\mathbf{u}_n\} \in S$ which converges to $\mathbf{x}$.

DEFINITION. A set S is open if for all $\mathbf{x} \in S$ there exists an open ball $B_r(\mathbf{x})$ such that $B_r(\mathbf{x}) \subset S$.

An example of a closed set is the closed interval $[0, 1] \subset \mathbb{R}$. An example of an open set is the open interval $(0, 1) \subset \mathbb{R}$. The complement of an open set is closed, and the complement of a closed set is open. The empty set is both open and closed, and so is $\mathbb{R}^n$. Given a set S and some point $\mathbf{b}$ outside of S we want to determine under what conditions there is a point $\hat{\mathbf{b}} \in S$ closest to $\mathbf{b}$. That is, such that $\|\hat{\mathbf{b}} - \mathbf{b}\| = d(\mathbf{b}, S)$ where $d(\mathbf{b}, S) = \inf_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{b}\|$. The quantity on the right of this definition is the greatest lower bound of the set of numbers $\|\mathbf{x} - \mathbf{b}\|$, and its existence is guaranteed by the properties of the real number system. What is not guaranteed in advance, and most be proved here, is the existence of an element $\hat{\mathbf{b}}$ which satisifes the

equality above. To see the problem, take $S = (0,1) \subset \mathbb{R}$ and $\mathbf{b} = 2$, then $d(\mathbf{b}, S) = 1$ yet there is no point $\hat{\mathbf{b}} \in (0,1)$ such that $\|\hat{\mathbf{b}} - 2\| = 1$.

THEOREM 1.1. *If S is a closed linear subspace of V then there exists $\hat{\mathbf{b}} \in S$ such that $\|\hat{\mathbf{b}} - \mathbf{b}\| = d(\mathbf{b}, S)$.*

PROOF. There exists a sequence of elements $\{\mathbf{u}_n\} \subset S$ such that $\|\mathbf{b} - \mathbf{u}_n\| \to d(\mathbf{b}, S)$ by definition of the greatest lower bound. We now show that this sequence is a Cauchy sequence.

From the parallelogram law we have

$$\left\|\frac{1}{2}(\mathbf{b} - \mathbf{u}_m)\right\|^2 + \left\|\frac{1}{2}(\mathbf{b} - \mathbf{u}_n)\right\|^2 = \frac{1}{2}\left\|\mathbf{b} - \frac{1}{2}(\mathbf{u}_n + \mathbf{u}_m)\right\|^2 + \frac{1}{4}\|\mathbf{u}_n - \mathbf{u}_m\|^2. \tag{1.2}$$

S is a vector space, therefore

$$\frac{1}{2}(\mathbf{u}_n + \mathbf{u}_m) \in S \Rightarrow \left\|\mathbf{b} - \frac{1}{2}(\mathbf{u}_n + \mathbf{u}_m)\right\|^2 \geq d^2.$$

Then since $\|\mathbf{b} - \mathbf{u}_n\| \to d(\mathbf{b}, S)$, we have:

$$\left\|\frac{1}{2}(\mathbf{b} - \mathbf{u}_n)\right\|^2 \to \frac{1}{4}d^2(\mathbf{b}, S).$$

From (1.2) above,

$$\|\mathbf{u}_n - \mathbf{u}_m\| \to 0,$$

and thus $\{\mathbf{u}_n\}$ is a Cauchy sequence by definition; our space is complete therefore this sequence converges to an element in this space, and S is closed, therefore the limit is in S. Finally,

$$\hat{\mathbf{b}} \in S \Rightarrow \|\hat{\mathbf{b}} - \mathbf{b}\| = d(\mathbf{b}, S).$$

■

We now wish to describe further the relation between $\mathbf{b}$ and $\hat{\mathbf{b}}$.

THEOREM 1.2. *Let S be a closed linear subspace of V, $\mathbf{x}$ an element of V not in S, and $\hat{\mathbf{b}}$ the element of S closest to $\mathbf{b}$. Then*

$$(\mathbf{x} - \hat{\mathbf{b}}, \mathbf{b} - \hat{\mathbf{b}}) \leq 0.$$

PROOF. Consider the vector $\theta(\mathbf{x} - \hat{\mathbf{b}}) - (\mathbf{b} - \hat{\mathbf{b}})$ where $0 < \theta \leq 1$. Because S is a vector space we have $\theta\mathbf{x} + (1-\theta)\hat{\mathbf{b}} \in S$, so that

$$\|\theta(\mathbf{x} - \hat{\mathbf{b}}) - (\mathbf{b} - \hat{\mathbf{b}})\|^2 = \|\theta\mathbf{x} + (1-\theta)\hat{\mathbf{b}} - \mathbf{b}\|^2 \geq d^2. \tag{1.3}$$

Note also that

$$\begin{aligned}\|\theta(\mathbf{x}-\hat{\mathbf{b}})-(\mathbf{b}-\hat{\mathbf{b}})\|^2 &= (\theta(\mathbf{x}-\hat{\mathbf{b}})-(\mathbf{b}-\hat{\mathbf{b}}),\theta(\mathbf{x}-\hat{\mathbf{b}})-(\mathbf{b}-\hat{\mathbf{b}})) \\ &= \theta^2(\mathbf{x}-\hat{\mathbf{b}},\mathbf{x}-\hat{\mathbf{b}})+(\mathbf{b}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}}) \\ &\quad -2\theta(\mathbf{x}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}}).\end{aligned}$$

However, by (1.3) we know that

$$\theta^2(\mathbf{x}-\hat{\mathbf{b}},\mathbf{x}-\hat{\mathbf{b}})+(\mathbf{b}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})-2\theta(\mathbf{x}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})\geq d^2.$$

By definition, $(\mathbf{b}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})=d^2$, therefore

$$\theta^2(\mathbf{x}-\hat{\mathbf{b}},\mathbf{x}-\hat{\mathbf{b}})-2\theta(\mathbf{x}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})\geq 0$$

and

$$\theta(\mathbf{x}-\hat{\mathbf{b}},\mathbf{x}-\hat{\mathbf{b}})-2(\mathbf{x}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})\geq 0$$

since $\theta>0$. By letting $\theta\to 0$, we obtain our result. ■

THEOREM 1.3. *$(\mathbf{b}-\hat{\mathbf{b}})$ is orthogonal to $\mathbf{x}$ for all $\mathbf{x}\in S$.*

PROOF. By theorem 1.2, $(\mathbf{x}-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})\leq 0$ for all $\mathbf{x}\in S$. Say $\mathbf{x}-\hat{\mathbf{b}}=\mathbf{w}$; we can find $\mathbf{x}'\in S$ such that $\mathbf{x}'-\hat{\mathbf{b}}=-\mathbf{w}$. Then:

$$(\mathbf{x}'-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})\leq 0$$

but

$$(\mathbf{x}'-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})\geq 0.$$

Therefore

$$(\mathbf{x}'-\hat{\mathbf{b}},\mathbf{b}-\hat{\mathbf{b}})=0.$$

Since $\mathbf{x}'-\hat{\mathbf{b}}$ is arbitrary in S, we are done. ■

COROLLARY 1.4. *If S is a closed linear subspace then $\hat{\mathbf{b}}$ is unique.*

PROOF. Let $\mathbf{b}=\hat{\mathbf{b}}+\mathbf{n}=\hat{\mathbf{b}}_1+\mathbf{n}_1$ Therefore:

$$\begin{aligned}\hat{\mathbf{b}}-\hat{\mathbf{b}}_1\in S &\Rightarrow (\hat{\mathbf{b}}-\hat{\mathbf{b}}_1,\mathbf{n}_1-\mathbf{n})=0 \\ &\Rightarrow (\hat{\mathbf{b}}-\hat{\mathbf{b}}_1,\hat{\mathbf{b}}-\hat{\mathbf{b}}_1)=0 \\ &\Rightarrow \hat{\mathbf{b}}=\hat{\mathbf{b}}_1.\end{aligned}$$

■

One can think of $\hat{\mathbf{b}}$ as the orthogonal projection of $\mathbf{b}$ on S, and write $\hat{\mathbf{b}}=P\mathbf{b}$, where the projection P is defined by the foregoing discussion.

We will now give a few applications of the above results.

EXAMPLE. Consider a matrix equation $A\mathbf{x} = \mathbf{b}$ where A is an $n \times m$ matrix and $n > m$. This kind of problem arises when one tries to fit a large set of data by a simple model. Assume the columns of A are linearly independent. Under what conditions does the system have a solution? To clarify ideas, consider the 3×2 case

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

Let A_1 denote the first column vector of A, A_2 the second column vector, etc. In the this case

$$A_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}, \quad A_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}.$$

If $A\mathbf{x} = \mathbf{b}$ has a solution, then one can express b as a linear combination of $A_1, A_2, \ldots A_m$, e.g., in the 3×2 case $x_1 A_1 + x_2 A_2 = b$. If $\mathbf{b}$ does not lie in the column space of A (the set of all all linear combinations of the columns of A), then the problem has no solution. It is often reasonable to replace the unsolvable problem by the solvable problem $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$ where $\hat{\mathbf{b}}$ is as close as possible to $\mathbf{b}$ and yet does lie in the column space of A. We know from the foregoing that the "best $\hat{\mathbf{b}}$" is such that $\mathbf{b} - \hat{\mathbf{b}}$ is orthogonal to the column space of A. This is enforced by the m equations:

$$(A_1, \hat{\mathbf{b}} - \mathbf{b}) = 0, \quad (A_2, \hat{\mathbf{b}} - \mathbf{b}) = 0, \quad \ldots \quad , (A_m, \hat{\mathbf{b}} - \mathbf{b}) = 0.$$

Since $\hat{\mathbf{b}} = A\hat{\mathbf{x}}$, we obtain the equation

$$A^T(A\hat{\mathbf{x}} - \mathbf{b}) = 0 \quad \Rightarrow \quad \hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}.$$

One application of the above is to "fit" a line to a set of points on the Euclidean plane. Given a set of points, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ which come from some experiment and that we believe would lie on a straight line if it were not for experimental error, what is the line that "best approximates" these points? We hope that if it were not for the errors, we would have $y_i = ax_i + b$ for all i, and for some a and b, so we seek to solve a system of equations

$$\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

EXAMPLE. Consider the system of equations given by $A\mathbf{x} = \mathbf{b}$ where A is an $n \times m$ matrix and $m < n$ (there are more unknowns

than equations). The system has infinitely many solutions. Suppose you want the solution of smallest norm; this problem arises when one tries to find the most likely solution to an underdetermined problem.

Before solving this problem we need some preliminaries:

DEFINITION. $S \subset V$ is an affine subspace if $S = \{\mathbf{y} : \mathbf{y} = \mathbf{x}+\mathbf{c}, \mathbf{c} \neq 0, \mathbf{x} \in X\}$ where X is a linear subspace of V. Note that S is not a linear subspace.

LEMMA 1.5. *If S is an affine subspace and $\mathbf{b}' \notin S$, then there exists $\hat{\mathbf{x}} \in X$ such that $d(\mathbf{b}', S) = \|\hat{\mathbf{x}} + \mathbf{c} - \mathbf{b}'\|$. Furthermore, $\hat{\mathbf{x}} - (\mathbf{b}' - \mathbf{c})$ is orthogonal to $\mathbf{x}$ for all $\mathbf{x} \in X$.(Note that here we use $\mathbf{b}'$ instead of $\mathbf{b}$, to avoid confusion with the system's RHS.)*

PROOF. We have $S = \{\mathbf{y} : \mathbf{y} = \mathbf{x} + \mathbf{c}, \mathbf{c} \neq 0, \mathbf{x} \in X\}$ where X is a closed linear subspace of V. Thus there exists $\mathbf{x}' \in X$ such that $d(\mathbf{b}', X) = d(\mathbf{x}', \mathbf{b}')$. Now

$$d(\mathbf{b}', S) = \inf_{\mathbf{y} \in S} \|\mathbf{y} - \mathbf{b}'\| = \inf_{\mathbf{x} \in X} \|\mathbf{x} + \mathbf{c} - \mathbf{b}'\|.$$

The latter occurs when $\mathbf{x} + \mathbf{c} = \mathbf{x}'$ and we denote this member of X as $\hat{\mathbf{x}}$, i.e., $\hat{\mathbf{x}} = \mathbf{x}' - \mathbf{c}$. Hence

$$d(\mathbf{b}', S) = d(\mathbf{x}', \mathbf{b}') = d(\hat{\mathbf{x}} + \mathbf{c}, \mathbf{b}') = \|\hat{\mathbf{x}} + \mathbf{c} - \mathbf{b}'\|.$$

Note that the distance between S and $\mathbf{b}'$ is the same as that between X and $\mathbf{b}'$. It follows from Theorem 1.3 that $\hat{\mathbf{x}} + \mathbf{c} - \mathbf{b}'$ is orthogonal to X. ■

From the proof above we see that $\hat{\mathbf{x}} + \mathbf{c}$ is the element of S closest to $\mathbf{b}'$. For the case $\mathbf{b}' = 0$ we find that $\hat{\mathbf{x}} + \mathbf{c}$ is orthogonal to X.

Now we return to the problem of finding the "smallest" solution of an underdetermined problem. Assume A has "maximal rank", i.e., m of the column vectors of A are linearly independent. We can write the solutions of the system as $\mathbf{x} = \mathbf{x}_0 + \mathbf{z}$ where $\mathbf{x}_0$ is a particular solution and $\mathbf{z}$ is the solution to the homogeneous system $A\mathbf{z} = 0$. So the solutions of the system $A\mathbf{x} = \mathbf{b}$ form an affine subspace. As a result, if we want to find the solution with the smallest norm, i.e., closest to the origin, we need to find the element of this affine subspace closest to $\mathbf{b}' = 0$. From the above we see that such an element must satisfy two properties. First, it has to be an element of the affine subspace, i.e., a solution to the system $A\mathbf{x} = \mathbf{b}$, and second, it has to be orthogonal to the linear subspace X, which now is the null space (the solutions of $A\mathbf{z} = 0$). For this purpose, consider $\mathbf{x}' = A^T(AA^T)^{-1}\mathbf{b}$; this vector lies in the affine subspace of the solutions of $A\mathbf{x} = \mathbf{b}$, as one can check by multiplying it by A; furthermore, it is orthogonal to every

vector in the space of solutions of $A\mathbf{z} = 0$ because $(A^T(AA^T)^{-1}\mathbf{b}, \mathbf{z}) = ((AA^T)^{-1}\mathbf{b}, A\mathbf{z}) = 0$. This is enough to make $\mathbf{x}'$ the unique solution of our problem.

2. Another Approach to Solving Least Squares Problems

The problem presented in the previous section, of finding an element in a closed linear space that is closest to a vector outside the space, lies in the framework of approximation theory where we are given a function (or a vector) and try to find an approximation to it as a linear combination of given functions (or vectors). This is done by requiring that the norm of the error between the given function and the approximation be minimized. In what follows we shall find coefficients for this optimal linear combination.

DEFINITION. Let S be an m-dimensional linear vector space. A collection of m vectors $\{\mathbf{u}_i\}_{i=1}^m$ belonging to S are linearly independent if and only if $\lambda_1\mathbf{u}_1 + \ldots + \lambda_m\mathbf{u}_m = 0$ implies $\lambda_1 = \lambda_2 = \ldots = \lambda_m = 0$.

DEFINITION. Let S be a linear vector space. A collection $\{\mathbf{u}_i\}_{i=1}^m$ of vectors belonging to S is called a basis of S if $\{\mathbf{u}_i\}$ are linearly independent and any vector in S can be written as a linear combination of them.

Note that the number of elements of a basis can be finite or infinite depending on the space.

THEOREM 1.6. *Let S be an m-dimensional linear space. Then any collection of m linearly independent vectors of S is a basis.*

DEFINITION. A set of vectors $\{\mathbf{e}_i\}_{i=1}^m$ is orthonormal if the vectors are mutually orthogonal and each has unit length, i.e., $(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

The set of all the linear combinations of the vectors $\{\mathbf{u}_i\}$ is called the span of $\{\mathbf{u}_i\}$ and is written as $\text{Span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$.

Suppose we are given a set of vectors $\{\mathbf{e}_i\}_{i=1}^m$ which are an orthonormal basis of S. If $\mathbf{b}$ is an element outside the space we want to find the element $\hat{\mathbf{b}} \in S$, where $\hat{\mathbf{b}} = \sum_{i=1}^m c_i\mathbf{e}_i$ such that $\|\mathbf{b} - \sum_{i=1}^m c_i\mathbf{e}_i\|$ is

minimized. Specifically we have:

$$\begin{aligned}\left\|\mathbf{b}-\sum_{i=1}^{m}c_i\mathbf{e}_i\right\|^2 &= \left(\mathbf{b}-\sum_{i=1}^{m}c_i\mathbf{e}_i\,,\mathbf{b}-\sum_{j=1}^{m}c_j\mathbf{e}_j\right)\\ &= (\mathbf{b},\mathbf{b})-2\sum_{i=1}^{m}c_i(\mathbf{b},\mathbf{e}_i)+\left(\sum_{i=1}^{m}c_i\mathbf{e}_i\,,\sum_{j=1}^{m}c_i\mathbf{e}_j\right)\\ &= (\mathbf{b},\mathbf{b})-2\sum_{i=1}^{m}c_i(\mathbf{b},\mathbf{e}_i)+\sum_{i,j=1}^{m}c_ic_j(\mathbf{e}_i,\mathbf{e}_j)\\ &= (\mathbf{b},\mathbf{b})-2\sum_{i=1}^{m}c_i(\mathbf{b},\mathbf{e}_i)+\sum_{i=1}^{m}c_i^2\\ &= \|\mathbf{b}\|^2-\sum_{i=1}^{m}(\mathbf{b},\mathbf{e}_i)^2+\sum_{i=1}^{m}(c_i-(\mathbf{b},\mathbf{e}_i))^2\end{aligned}$$

where we have used the orthonormality of the $\mathbf{e}_i$ to simplify the expression. As is readily seen, the norm of the error is a minimum when $c_i=(\mathbf{b},\mathbf{e}_i), i=1,m$, so that $\hat{\mathbf{b}}$ is the projection of $\mathbf{b}$ onto S. It is easy to check that $\mathbf{b}-\hat{\mathbf{b}}$ is orthogonal to any element in S. Also, we see that the following inequality, called Bessel's inequality, holds

$$\sum_{i=1}^{m}(\mathbf{b},\mathbf{e}_i)^2\le\|\mathbf{b}\|^2.$$

When the basis is not orthonormal, steps similar to the above yield:

$$\begin{aligned}\|\mathbf{b}-\sum_{i=1}^{m}c_i\mathbf{g}_i\|^2 &= (\mathbf{b}-\sum_{i=1}^{m}c_i\mathbf{g}_i,\mathbf{b}-\sum_{j=1}^{m}c_j\mathbf{g}_j)\\ &= (\mathbf{b},\mathbf{b})-2\sum_{i=1}^{m}c_i(\mathbf{b},\mathbf{g}_i)+(\sum_{i=1}^{m}c_i\mathbf{g}_i,\sum_{j=1}^{m}c_j\mathbf{g}_j)\\ &= (\mathbf{b},\mathbf{b})-2\sum_{i=1}^{m}c_i(\mathbf{b},\mathbf{g}_i)+\sum_{i,j=1}^{m}c_ic_j(\mathbf{g}_i,\mathbf{g}_j).\end{aligned}$$

If we differentiate the last expression with respect to c_i and set the derivatives equal to zero we get

$$G\mathbf{c}=\mathbf{r}$$

where G is the matrix with entries $g_{ij}=(\mathbf{g}_i,\mathbf{g}_j)$, $\mathbf{c}=(c_1,\ldots,c_m)^T$ and $\mathbf{r}=((\mathbf{g}_1,\mathbf{b}),\ldots,(\mathbf{g}_m,\mathbf{b}))^T$. This system can be ill-conditioned so that its numerical solution presents a problem. The question that arises is

how to find, given a set of vectors, a new set that is orthonormal. This is done through the Gram-Schmidt process which we now describe.

Let $\{\mathbf{u}_i\}_{i=1}^m$ be a basis of a linear subspace. The following algorithm will give an orthonomal set of vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ such that $\text{Span}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\} = \text{Span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$.

(1) Normalize $\mathbf{u}_1$, i.e., let $\mathbf{e}_1 = \mathbf{u}_1/\|\mathbf{u}_1\|$.
(2) We want a vector $\mathbf{e}_2$ that is orthonormal to $\mathbf{e}_1$. In other words we look for a vector $\mathbf{e}_2$ satisfying $(\mathbf{e}_2, \mathbf{e}_1) = 0$ and $\|\mathbf{e}_2\| = 1$. Take $\mathbf{e}_2 = \mathbf{u}_2 - (\mathbf{u}_2, \mathbf{e}_1)\mathbf{e}_1$ and then normalize.
(3) In general, $\mathbf{e}_j$ is found recursively by taking

$$\mathbf{e}_j = \mathbf{u}_j - \sum_{i=1}^{j-1} (\mathbf{u}_j, \mathbf{e}_i)\mathbf{e}_i$$

and normalizing.

EXAMPLE. Let $f(x) \in C[0, 2\pi]$ with inner product

$$(f, g) = \int_0^{2\pi} f(x)g(x)dx.$$

What is the closest polynomial of degree 7 to $f(x)$ (i.e., a polynomial P_7 such that $\int_0^{2\pi}(f(x) - P_7(x))^2 dx$ is minimized over all polynomials of degree ≤ 7)? Note that the "best" P_7 does exist because the collection of polynomials of degree less than or equal to 7 is a closed linear subspace. Begin by finding an orthonormal basis $\{\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_7\}$: Take $\mathbf{b}_0 = c_0$ where $\int_0^{2\pi} c_0^2 dx = 1 \Rightarrow c_0 = 1/\sqrt{2\pi}$. Next we have $\mathbf{b}_1 = c_0 + c_1 x$ where $\int_0^{2\pi} \mathbf{b}_1^2 dx = 1$, etc.

The Gram-Schmidt process can be implemented numerically very efficiently. The process can be rewritten in matrix form as:

$$\begin{aligned}
\mathbf{e}_1 &= a_{11}\mathbf{u}_1 \\
\mathbf{e}_2 &= a_{12}\mathbf{u}_1 + a_{22}\mathbf{u}_2 \\
&\vdots \\
\mathbf{e}_m &= a_{1m}\mathbf{u}_1 + a_{2m}\mathbf{u}_2 + \ldots + a_{mm}\mathbf{u}_m
\end{aligned}$$

with $(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}$. The solution of this system is equivalent to finding $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$, such that the following holds

$$\begin{aligned}
\mathbf{u}_1 &= b_{11}\mathbf{e}_1 \\
\mathbf{u}_2 &= b_{12}\mathbf{e}_1 + b_{22}\mathbf{e}_2 \\
&\vdots \\
\mathbf{u}_m &= b_{1m}\mathbf{e}_1 + b_{2m}\mathbf{e}_2 + \ldots + b_{mm}\mathbf{e}_m
\end{aligned}$$

i.e., what we want to do is decompose the matrix U with columns $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ into a product of two matrices Q and R, where Q has as columns the orthonormal vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ and R is the matrix

$$R = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ 0 & b_{22} & \ldots & b_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & b_{mm} \end{bmatrix}.$$

This is the well-known QR decomposition and there exist efficient ways to implement it.

3. Fourier Series

Let $L_2[0, 2\pi]$ be the space of square integrable functions in $[0, 2\pi]$, i.e., such that $\int_0^{2\pi} f^2 dx < \infty$. Define the inner product of two functions f and g belonging to this space as $(f, g) = \int_0^{2\pi} fg dx$ and the corresponding norm $\|f\| = \sqrt{(f, f)}$. The Fourier series of a function $f(x)$ in this space is defined as

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx) \tag{1.4}$$

where

$$\begin{aligned}
a_0 &= \frac{1}{2\pi} \int_0^{2\pi} f(x) dx, \\
a_n &= \frac{1}{\pi} \int_0^{2\pi} \cos(nx) f(x) dx, \\
b_n &= \frac{1}{\pi} \int_0^{2\pi} \sin(nx) f(x) dx.
\end{aligned}$$

Alternatively, consider the set

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos(nx), \frac{1}{\sqrt{\pi}} \sin(nx), \ldots \right\}, \quad n = 1, 2, \ldots.$$

This set is orthonormal in $[0, 2\pi]$ and the Fourier series (1.4) can be rewritten as

$$f(x) = \frac{\tilde{a}_0}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \frac{\tilde{a}_n}{\sqrt{\pi}} \cos(nx) + \sum_{n=1}^{\infty} \frac{\tilde{b}_n}{\sqrt{\pi}} \sin(nx). \tag{1.5}$$

For any function in L_2 the series (1.5) converges in the L_2 norm, i.e., let

$$S_0 = \frac{\tilde{a}_0}{\sqrt{2\pi}}, \quad S_n = \frac{\tilde{a}_0}{\sqrt{2\pi}} + \sum_{m=1}^{n} \frac{\tilde{a}_m}{\sqrt{\pi}} \cos mx + \sum_{m=1}^{n} \frac{\tilde{b}_m}{\sqrt{\pi}} \sin mx \text{ for } n \geq 1$$

with

$$\begin{aligned}
\tilde{a}_0 &= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(x)dx, \\
\tilde{a}_n &= \frac{1}{\sqrt{\pi}} \int_0^{2\pi} \cos(nx) f(x)dx, \\
\tilde{b}_n &= \frac{1}{\sqrt{\pi}} \int_0^{2\pi} \sin(nx) f(x)dx.
\end{aligned}$$

Then we have $\|S_n - f\| \to 0$ as $n \to \infty$.

For any finite truncation of the series (1.5) we have

$$\tilde{a}_0^2 + \sum_{i=1}^{n} \left(\tilde{a}_i^2 + \tilde{b}_i^2 \right) \leq \|f\|^2. \tag{1.6}$$

This is the Bessel inequality which becomes an equality (Parseval equality) as $n \to \infty$.

The above series (1.5) can be rewritten in complex notation. Recall that

$$\cos(kx) = \frac{e^{ikx} + e^{-ikx}}{2}, \quad \sin(kx) = \frac{e^{ikx} - e^{-ikx}}{2i}. \tag{1.7}$$

After substitution of (1.7) into (1.5) and collection of terms the Fourier series becomes

$$f(x) = \sum_{k=-\infty}^{\infty} \frac{c_k}{\sqrt{2\pi}} e^{ikx}$$

where f is now complex. (Note that f will be real if for $k \geq 0$ we have $c_{-k} = \overline{c_k}$.) Consider a vector space with complex scalars and introduce an inner product that satisfy the axioms (1.1) and define the norm $\|u\| = \sqrt{(u, u)}$. For the special case where the inner product is given by

$$(u, v) = \int_0^{2\pi} u(x)\bar{v}(x)dx,$$

the functions $(2\pi)^{-1/2}\, e^{ikx}$ with $k = 0, \pm 1, \pm 2, \ldots$ form an orthonormal set with respect to this norm. Then the complex Fourier series of a complex function $f(x)$ is written as

$$f(x) = \sum_{k=-\infty}^{\infty} \tilde{c}_k \frac{1}{\sqrt{2\pi}} e^{ikx}, \quad c_k = \left(f(x), \frac{e^{-ikx}}{\sqrt{2\pi}} \right).$$

Let $f(x)$ and $g(x)$ be two functions with Fourier series given respectively by

$$f(x) = \sum_{k=-\infty}^{\infty} \frac{a_k}{\sqrt{2\pi}} e^{ikx},$$
$$g(x) = \sum_{k=-\infty}^{\infty} \frac{b_k}{\sqrt{2\pi}} e^{ikx}.$$

Then for their inner product we have

$$(f, g) = \int_0^{2\pi} f(x)\bar{g}(x)dx = \int_0^{2\pi} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \frac{a_k \bar{b}_l}{2\pi} e^{i(k-l)x} = \sum_{k=-\infty}^{\infty} a_k \bar{b}_k$$

and their product we have

$$f(x)g(x) = \sum_{k=-\infty}^{\infty} \frac{c_k}{\sqrt{2\pi}} e^{ikx}$$

where

$$\begin{aligned} c_k &= \int_0^{2\pi} \left(\sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \frac{a_n b_m}{2\pi} e^{i(n+m)x} \right) \frac{e^{-ikx}}{\sqrt{2\pi}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_n b_m \delta(n + m - k) \\ &= \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} a_n b_{k-n} = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} a_{k-n} b_n. \end{aligned}$$

4. Fourier Transform

Consider the space of periodic functions defined on the interval $[-\tau/2, \tau/2]$. The functions $\tau^{-1/2} \exp(2\pi ikx/\tau)$ are an orthonormal basis for this space. For a function $f(x)$ in this space we have

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e_k(x), \quad c_k = (f, \overline{e_k}(x))$$

where

$$e_k(x) = \frac{\exp(2\pi ikx/\tau)}{\sqrt{\tau}}$$

and

$$(f, \overline{e_k}) = \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(x)\overline{e_k}(x)dx.$$

Substituting the expression for the coefficient in the series we have

$$\begin{aligned} f(x) &= \sum_{k=-\infty}^{\infty} \left(\int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(s) \frac{\exp(2\pi iks/\tau)}{\sqrt{\tau}} ds \right) \frac{\exp(-2\pi ikx/\tau)}{\sqrt{\tau}} \\ &= \sum_{k=-\infty}^{\infty} \frac{1}{\tau} \left(\int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(s) \exp(2\pi iks/\tau) ds \right) \exp(-2\pi ikx/\tau). \end{aligned}$$

Define

$$\hat{f}(l) = \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(s) \exp(-ils) ds.$$

Then the quantity in parantheses above becomes $\hat{f}(l = 2\pi k/\tau)$ and we have

$$f(x) = \sum_{k=-\infty}^{\infty} \frac{1}{\tau} \hat{f}(2\pi k/\tau) \exp(2\pi ikx/\tau). \tag{1.8}$$

Pick τ large and assume that the function f tends to zero at $\pm\infty$ fast enough so that $\hat{f}$ is well defined and that the limit $\tau \to \infty$ is well defined. Write $\Delta = 1/\tau$. From (1.8) we have

$$f(x) = \sum_{k=-\infty}^{\infty} \Delta \hat{f}(2\pi k\Delta) \exp(2\pi ik\Delta x).$$

As $\Delta \to 0$ this becomes

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(2\pi t) \exp(2\pi itx) dt$$

where we have replaced k by the continuous variable t. By the change of variables $2\pi t = l$ this becomes

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(l) \exp(ilx) dl.$$

Collecting results we have

$$\begin{aligned} \hat{f}(l) &= \int_{-\infty}^{\infty} f(s) \exp(-ils) ds, \\ f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(l) \exp(ilx) dl. \end{aligned}$$

The last two expressions are the Fourier transform and the inverse Fourier transform respectively. There is no universal agreement on where the quantity 2π that accompanies the Fourier transform should be. It can be split between the Fourier transform and its inverse as long as the product remains 2π. In what follows we use the splitting

$$\hat{f}(l) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(s) \exp(-ils) ds,$$
$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(l) \exp(ilx) dl.$$

Instead of $L_2[0, 2\pi]$, now our space of functions is $L_2(\mathbb{R})$, i.e., the space of square integrable functions on the real line.

5. Properties of the Fourier Transform

Consider two functions $u(x)$ and $v(x)$ with Fourier series given respectively by $\sum a_k \exp(ikx)$ and $\sum b_k \exp(ikx)$. Then as we saw above the Fourier coefficients for their product are

$$c_k = \frac{1}{\sqrt{2\pi}} \sum_{k'=-\infty}^{\infty} a_{k'} b_{k-k'}.$$

This property carries over to the case of the Fourier transform, so for two functions f and g with Fourier transforms $\hat{f}$ and $\hat{g}$, we have

$$\begin{aligned} \widehat{fg} &= \int_{-\infty}^{\infty} \hat{f}(k') \hat{g}(k - k') dk' \\ &= \int_{-\infty}^{\infty} \hat{f}(k - k') \hat{g}(k') dk \\ &= \hat{f} * \hat{g} \end{aligned}$$

where $*$ stands for "convolution." This means that the Fourier transform of a product of two functions equals the convolution of the Fourier transforms of the two functions.

Another useful property of the Fourier transform concerns the Fourier transform of the convolution of two functions. Assuming f and g are bounded, continuous and integrable the following result holds for their

convolution $h(x) = f(x) * g(x)$:

$$\begin{aligned}\hat{h} = \widehat{(f*g)} &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(\xi)g(x-\xi)d\xi\right)e^{-ikx}dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi)g(y)e^{-iky}e^{ik\xi}d\xi dy \\ &= \int_{-\infty}^{\infty} f(\xi)e^{-ik\xi}d\xi \int_{-\infty}^{\infty} g(y)e^{-iky}dy \\ &= \hat{f}\hat{g}.\end{aligned}$$

Thus, we have proved that the Fourier transform of a convolution of two functions is the product of the Fourier transforms of the functions.

In addition, Parseval's equality carries over to the Fourier transform and we have $\|f\|^2 = \|\hat{f}\|^2$ where $\|\cdot\|$ is the L_2 norm on $\mathbb{R}$. We also have the results

$$\begin{aligned}(f+g, f+g) &= (\hat{f}+\hat{g}, \hat{f}+\hat{g}) \\ \|f\|^2 + \|g\|^2 + (g,f) + (f,g) &= \|\hat{f}\|^2 + \|\hat{g}\|^2 + (\hat{g},\hat{f}) + (\hat{f},\hat{g}) \\ Re(f,g) &= Re(\hat{f},\hat{g}).\end{aligned}$$

Futhermore, consider a function f and its Fourier transform $\hat{f}$. Then for the transform of the function $f(x/a)$ we have

$$\widehat{f\left(\frac{x}{a}\right)} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f\left(\frac{x}{a}\right)e^{-ikx}dx.$$

By the change of variables $y = x/a$ we have

$$\begin{aligned}\widehat{f\left(\frac{x}{a}\right)} &= \frac{a}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(y)\exp(-iaky)dy \\ &= a\hat{f}(ak).\end{aligned}$$

Finally, consider the function $f(x) = \exp(-x^2/2t)$ where t is a parameter. For its Fourier transform we have

$$\begin{aligned}\hat{f}(k) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{x^2}{2t}\right)\exp(-ikx)dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\left(\frac{x^2}{2t}+ikx\right)\right]dx.\end{aligned}$$

By completing the square in the exponent we get

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x}{\sqrt{2t}} + ik\sqrt{\frac{t}{2}} \right)^2 - \frac{tk^2}{2} \right] dx$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-tk^2/2\right) \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x}{\sqrt{2t}} + ik\sqrt{\frac{t}{2}} \right)^2 \right] dx. \quad (1.9)$$

The integral in the last expression can be evaluated by a change of variables, but we have to justify that such a change of variables is legitimate. To do that we quote a result from complex analysis.

LEMMA 1.7. *Let $\phi(z)$ be an analytic function in the strip $|y| < b$ and suppose that $\phi(z)$ satisfies the inequality $|\phi(x+iy)| \leq \Phi(x)$ in the strip where $\Phi(x) \geq 0$ is a function such that $\lim_{|x|\to\infty} \Phi(x) = 0$ and $\int_{-\infty}^{\infty} \Phi(x)dx < \infty$. Then the value of the integral $\int_{-\infty}^{\infty} \phi(x+iy)dx$ is independent of the point $y \in (-b, b)$.*

The integrand in (1.9) satisfies the hypotheses of the lemma and so we are allowed to perform the change of variables

$$y = \frac{x}{\sqrt{2t}} + ik\sqrt{\frac{t}{2}}.$$

Thus (1.9) becomes

$$\begin{aligned} \hat{f}(k) &= \frac{1}{\sqrt{2\pi}} \exp(-tk^2/2) \int_{-\infty}^{\infty} \exp(-y^2)\sqrt{2t}dy \\ &= \frac{1}{\sqrt{2\pi}} \exp(-tk^2/2)\sqrt{2t\pi} \\ &= \sqrt{t}\, \exp(-tk^2/2). \end{aligned}$$

By setting $t = 1$ we see in particular that the function $f(x) = \exp(-x^2/2)$ is invariant under the Fourier transform.

6. References

1. H. Dym and H. McKean, *Fourier Series and Integrals*, Academic Press, NY, 1972.
2. G. Golub and F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1983.
3. A. Kolmogorov and S. Fomin, *Elements of the Theory of Functions and Real Analysis*,Dover, 2000.
4. P. Lax, *Linear Algebra*, Wiley, NY, 1997.